DOCUMENT RESUME

ED 214 950 TM 820 091

AUTHOR Tyler, Ralph W.; White, Sheldon H.
TITLE Testing, Teaching and Learning: Chairmen's Report of a Conference on Research on Testing (August 17-26, 1978).
INSTITUTION National Inst. of Education (DHEW), Washington, D.C.; Office of the Assistant Secretary for Education (DHEW), Washington, D.C.
PUB DATE Oct 79
NOTE 42p.; For related document see ED 181 080.
AVAILABLE FROM Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402 (631-367-2858).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Culture Fair Tests; *Educational Improvement; *Educational Testing; Elementary Secondary Education; *Testing Problems; *Test Use
IDENTIFIERS Test Curriculum Overlap

ABSTRACT

Four broad issues of educational tests use are presented with their major criticisms: tests meant to hold educators and school systems accountable have limited value, and when used to make decisions concerning individual students tests do not fully reflect the cultural backgrounds of minority students. Testing to evaluate educational innovations and experimental projects is criticized as being too narrow in scope for fair evaluation. Test use to guide teachers is seen as exercising a limiting effect in the classroom. The first recommendation is to develop equivalent tests better tailored to cultural background. A further suggestion is to better fit testing to educational objectives by increased research on criterion-referenced skills tests; application of information-handling technology; and by clarification of basic testing concepts for educators, parents and policy makers. Appropriate test use is discussed. Merging testing with teaching is recommended, using cognitive science analyses, interactive automated teaching-testing and involvement of teachers and scholars in testmaking. Expanded research and development of testing-learning models are suggested within natural classroom situations. A partial list of new test strategies, references, comments and an appendix of subject papers are given. (CM)

TESTING, TEACHING AND LEARNING

Chairmen's Report of a Conference on Research on Testing

The National Institute of Education
U.S. Department of Health, Education and Welfare
Washington, D.C. 20208

CONFERENCE PARTICIPANTS

GREGORY R. ANRIG, Commissioner of Education, State of Massachusetts
CARL BEREITER, Dept. of Applied Psychology, Ontario Institute for Studies in Education
JOHN SEELY BROWN, Xerox Palo Alto Research Laboratories
SYLVIA CHARP, Instructional Systems Divn., School District of Philadelphia
ALLAN COLLINS, Center for the Study of Reading, Bolt Beranek & Newman
ANN COOK, Community Resources Institute, New York, N.Y.
ALONZO A. CRIM, Superintendent, Atlanta Public Schools
PARKER DAMON, Principal, McCarthy-Towne School, Acton, Mass.
ROBERT B. DAVIS, Curriculum Laboratory, University of Illinois, Urbana
LAWRENCE T. FRASE, Bell Laboratories, Piscataway, N.J.
NORMAN O. FREDERIKSEN, Division of Psychological Studies, Educational Testing Service
ANDREW M. GLEASON, Dept. of Mathematics, Harvard University
DONALD ROSS GREEN, Director of Research, CTB/McGraw Hill, Monterey, Calif.
JAMES G. GREENO, Learning R&D Center, University of Pittsburgh
BARBARA HAYES-ROTH, Rand Corporation, Santa Monica, Calif.
LEON A. HENKIN, Dept. of Mathematics, University of California, Berkeley
ASA G. HILLIARD, School of Education, San Francisco State University
C. DALTON JONES, Dept. of Psychology, Univ. of Massachusetts, Amherst
DAVID KLAHR, Dept. of Psychology, Carnegie-Mellon University
BERTRAM L. KOSLIN, Touchstone Applied Science Associates, Elmsford, N.Y.
GEORGE A. MILLER, Dept. of Psychology, Princeton University
AMADO M. PADILLA, Dept. of Psychology, Univ. of California, Los Angeles
ANDREW C. PORTER, Institute for Research on Teaching, Michigan State Univ.
FRANCES QUINTO, National Education Association, Washington, D.C.
MARIA RAMIREZ, Asst. Commissioner, N.Y. State Education Department
JUDAH L. SCHWARTZ, Div. for Study and Research in Education, Massachusetts Institute of Technology
SUSAN S. STODOLSKY, Depts. of Education and Human Development, University of Chicago
ROSS TAYLOR, Mathematics Consultant K-12, Minneapolis Public Schools
RALPH W. TYLER, Senior Consultant, Science Research Associates
SHELDON H. WHITE, Dept. of Psychology & Social Relations, Harvard University
JERROLD R. ZACHARIAS, Education Development Center

TESTING, TEACHING AND LEARNING

Chairmen's Report
of a
Conference on
Research on Testing

August 17-26, 1978

Ralph W. Tyler
and
Sheldon H. White
Chairmen

U.S. Department of Health, Education, and Welfare
Patricia Roberts Harris, Secretary
Mary F. Berry, Assistant Secretary for Education

National Institute of Education
Michael Timpane, Acting Director

October 1979

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402

Points of view or opinions expressed in this report are not necessarily those of the National Institute of Education or the Department of Health, Education, and Welfare.

FOREWORD

The National Institute of Education (NIE), the primary federal agency for educational research and development, sponsored the conference reported on here with the aim of fostering research into better means of assessing what students have learned and the difficulties they are having in learning. Such assessments are central to improving the practice of education and promoting educational equity—two of the Institute's main goals. The conference, held in August 1978, was a response to the widespread interest in research on testing—interest made clear by participants in a more broadly representative Conference on Achievement Testing and Basic Skills, convened by the Secretary of Health, Education, and Welfare earlier in 1978.

The discussion of issues and opportunities in testing and the research agendas in this report are proving very useful to NIE in developing a research agenda in testing as part of its Program on Teaching and Learning. We believe they will also be interesting and useful to a variety of public and private organizations and individuals concerned with improving education, with testing, and with the development of tests.

Michael Timpane, Acting Director
National Institute of Education

CHAIRMEN'S PREFACE

Until the late 1960s, criticism of educational testing was largely confined to academic debate. Since then, testing has become a center of public controversy.

Minority groups object that published tests are unfair to their children. Many teachers, parents, and school administrators find existing tests unsatisfactory as indicators of a student's educational progress. The National Education Association has called for a moratorium on testing.

In March 1978, Secretary of Health, Education, and Welfare, Joseph A. Califano, Jr., convened the three-day National Conference on Achievement Testing and Basic Skills. Participants were chosen to be broadly representative of current concerns. One goal of the conference was to find ways in which HEW could help states and localities to use tests more effectively. Another was to open a national discussion about the reasonable use of tests to improve the quality of elementary and secondary education. Among its recommendations, the conference called for expansion in scope of Federal support of research on learning and on the uses of testing.

In August 1978, responding to this recommendation, the National Institute of Education, HEW's educational research agency, sponsored a ten-day conference on research on testing, planned by four members of the NIE staff. Sylvia Scribner, then Associate Director for Teaching and Learning, John M. Mays and Arthur S. Melmed in the Office of the Director, and Jeffry Schiller, Assistant Director for Testing, Assessment, and Evaluation.

The thirty-one participants at the conference comprised persons concerned with teaching and educational administration and policy, with research and development relevant to education, with educational testing, with various areas of educational content, and with information-handling technology.

Many of the points made at the first conference were also strongly asserted at the second conference:

- A national achievement test is no more the answer to improving educational quality than is a national curriculum.
- Beware of tests that are hastily constructed to meet demands that are hastily made.

- Everyone needs better information on what tests can and cannot do—test users, test takers (including their parents), and test makers.
- Current testing procedures are unfair to particular minority groups.
- Current testing procedures are not helpful to teachers or students in their day-to-day efforts to teach and learn.
- Tests are not the whole story in assessing educational progress, and ways must be found to make better uses of qualitative information from other sources.

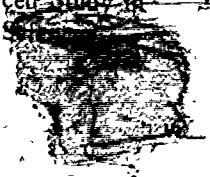
How, then, is research to help improve our use of tests? The general strategy of the conference was to address both short-term realities and long-term possibilities. As might be expected, the scope of the discussion was broad, ranging from present-day problems to test usage in schools to the state-of-the-art of computerized testing and teaching. Particular emphasis was placed on exploration of the implications for testing of three new areas of progress. First, recent advances in cognitive science which provide new understanding of human intellectual processes and allow better design of instruction and testing and better fitting of them to cultural background. Second, progress in information-handling technology, which is in a period of revolutionary increase in capability and lowering cost, allowing greatly increased breadth and individualization of instruction and testing. Third, the new sense of what is achievable in education that has grown out of collaboration between teachers and scholars in the last two decades.

The conference began with presentations of invited papers, which are listed in the Appendix. The conference then divided into three working committees, each reflecting the same diversity of interests and backgrounds as the conference as a whole. The working committees met separately, reporting to each other periodically at plenary sessions, and presented final reports on the last day of the conference. This Chairmen's Report, the papers presented at the conference, the reports of the working committees, and, as an Appendix, the report of the National Conference on Achievement Testing and Basic Skills are published in a separate volume *Testing, Teaching and Learning: Report of a Conference on Research on Testing*, which is for sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.

This Chairmen's Report, prepared with the help of Joseph Turner, sets forth from our perspective, the deliberations and recommendations of the conference. All participants have reviewed this document and their comments are incorporated in it or are appended.

Ralph W. Tyler
Director, Emeritus
Center for Advanced Study in
the Behavioral Sciences

Sheldon H. White
Dept. of Psychology
Harvard University



Contents

SUMMARY

The Issues Concerning Tests

Broadly speaking, educational tests are used in American education today to achieve four purposes:

1. Holding teachers, schools, and school systems accountable.
2. Making decisions concerning individual students.
3. Evaluating educational innovations and experimental projects.
4. Providing guidance to teachers in the classroom.

The major current criticisms of educational tests are the following.

1. Tests do not reflect the full range of student cultural backgrounds and thus lead to decisions that are unfair to minority students.
2. Tests have only limited value for holding educators accountable.
3. Tests exercise a limiting effect on classroom teaching.
4. Tests are too narrow in scope to provide for fair evaluation of new approaches to teaching.

Recommendations

The conference participants believe that testing, and thereby education, can be substantially improved through a variety of research and development and other related activities responding to the issues above. Recommendation A responds to the first criticism. Recommendation B responds to the last three criticisms, which arise in a considerable part from a lack of fit between testing and educational objectives. Recommendation C looks toward a system in the future in which testing is made more effective and helpful by being merged into the teaching process.

A. Better Fitting of Testing to Cultural Background of Students

•Development of equivalent culture-specific tests in which the difficulty level and knowledge domains are the same, but the language and illustrative material are tailored to the background knowledge of particular groups.

B. Better Fitting of Testing to Educational Objectives

Criterion-referenced Testing

Increasing research and development on the theory and practice of criterion-referenced tests, which are specifically designed to determine what a student knows and can do in a specified domain of educational objectives.

Information-handling Technology

Application to testing of the increasingly available low-cost information-handling technology in such areas as:

Computer-based Item Pools

Creation of computer-based pools of test items by educational centers and test publishers which would allow teachers, schools, and school systems to create tests closely matched to their educational objectives.

Testing Problem-solving Processes

Development of means of testing in one of the principal aims of education—the ability to address more complex problems of the sort encountered in work and personal life—now made possible by new low-cost computer capability.

Tailored Testing

Development of more tests in which the computer matches the level of difficulty of items to the capability demonstrated by students and thus provides better and quicker assessments.

Providing Better Information on Tests and Testing

Developing and making available the information on tests and testing that educators, parents, policymakers, and the public need to know in fitting testing to educational objectives:

Clarifying Basic Conceptions of Testing

Taking a new look at what we can and cannot expect from tests in the light of present understanding of the complex nature of human knowledge—and making the results easily available to interested parties.

Appropriate Use of Tests in Education

A study of the use of tests in American education documenting the typical uses of tests and investigating the effects of testing on teaching, on the education of individuals, on educational innovation, and on budgetary and other decisions—to provide much firmer information on the suitability of various testing practices than is now possible.

Information on Published Tests

Formation of privately supported consumer groups as sources of information about the availability and capabilities of various tests, and provision for fully informative disclosure of test items after administration.

C. Combining Testing with Teaching

Sustained research and development over the next decade or more toward a system in which testing is merged into the teaching process and provides timely and rich feedback to the student, the teacher, and other appropriate parties. Four elements are seen as essential to this R&D program:

Cognitive Science

Cognitive science, an interdisciplinary movement comprising psychologists, computer scientists, philosophers, linguists, anthropologists, and educators, is providing powerful new understanding of the processes in learning and performing the intellectual tasks to be tested.

Interactive Teaching-Testing and Technology

The rapidly increasing capacity and decreasing cost of electronic information-handling technology make feasible branching, exploratory interplay with the learner, of which testing can be a natural part.

Subject Matter

Involvement of knowledgeable teachers and scholars in test making, important for any test, is essential in the development of teaching-testing systems and tests of problem-solving processes.

The Natural Classroom Situation

The teaching-testing systems envisioned are intended as aids to teachers rather than substitutes for them and thus must be developed making full use of what is known about the way instruction and instructional guidance take place at the grade level and in the subject area in question and about the successes and failures of previous innovations.

THE ISSUES CONCERNING TESTS

The theory that has, explicitly or implicitly, dominated most thinking about tests is that they provide measures, analogous to length and weight, that can be used to arrange children and what they have learned, in a simple linear order. This theory supported initial research on testing, but it is obviously inadequate to our present conceptions of the complex nature of human knowledge.

George A. Miller

The widespread use of tests suggests that they fill the needs of many people. Who uses tests? When? For what purposes? Although there is now much criticism of testing, a careful observer will note that many critics examine specific issues rather than find fault across the board and that different critics pose different objections. We consider first the major uses of current tests and then the chief criticisms leveled at them.

Major Uses of Current Tests

Broadly speaking, educational tests are used in American education today to achieve four purposes: accountability, selection, evaluation of experimental projects, and instructional guidance.

Tests are used to hold teachers, schools, and school systems accountable.

Many principals, superintendents, and other educational authorities use test scores, particularly scores on achievement tests, as a rough gauge of the adequacy of the performance of a teacher, a school, or larger administrative unit. Parents, voters, and legislators also use such information in judging schools and school systems. The results of a test are taken to indicate the amount of learning accomplished by the average student in a classroom or larger unit. Educational authorities may then decide that the teaching offered is either adequate, inferior, or superior.

Tests are used to make decisions concerning individual students.

Educational authorities use the same tests when placing individual

students in special programs and classes and in counselling them on plans for future education and careers. Procedures vary from one system to another, but combinations of ability and achievement tests are generally used for these purposes. Authorities may decide to assign a pupil in the early grades to a slow-paced program if both ability and achievement scores are low, or to a program for the gifted if both are high. A low score in reading achievement coupled with a high ability score raises the possibility of a reading or learning disability; more specialized diagnostic tests may then be used.

College admissions officers depend heavily on student scores on the Scholastic Aptitude Test (SAT) and similar tests in considering applicants. Some colleges make the achievement of a score above some cutoff point a necessary but not sufficient condition for admission. School grades, letters of recommendation, and interviews are significant, but exactly which factors are considered and how they are weighed varies from college to college and from time to time.

Tests are used to evaluate educational innovations and experimental projects.

Government agencies, private foundations, and school systems sponsor experimental projects in American schools and seek to evaluate these projects through use of standardized achievements tests. A recent wave of experimental projects was the curriculum reform movement in science and mathematics, which began in the 1950s. Another, larger wave came in the 1960s when widespread efforts were made to improve the education of children from backgrounds of poverty and discrimination. Evaluators of experimental projects continue to wrestle with the task of matching tests to project objectives. In some cases, experimenters have found available tests unsuited to their projects and have developed new ones.

Tests are used to provide guidance to teachers in the classroom.

Test makers and the educators who select published tests for use in schools exert influence directly and indirectly on teachers in the classrooms. Direct influence is exerted by the choice of subtests to measure the strengths and weaknesses of students in particular component skills. Ideally, a teacher studying the scores could determine where to invest teaching time with particular students or classes. The indirect influence results when tests are used for accountability. If teachers are to be rated by administrators on the basis of their students' scores, prudent teachers will emphasize in their teaching the topics emphasized in the tests.

Major Criticisms of Current Tests

The major present-day criticisms of educational testing are the following:

Tests do not reflect the full range of student cultural backgrounds and thus lead to decisions that are unfair to minority students.

The underlying logic of standardized testing requires that a given test performance must have the same meaning for all children or groups of children being assessed or compared. We would not have much confidence in a thermometer if its operation varied with certain background characteristics of patients that were irrelevant to their physical condition. Similarly, our confidence that educational assessments are accurate and selection decisions fair rests on the condition that test performance is not affected in any systematic way by irrelevant background characteristics of children—whether they are boys or girls, black or white, or of Hispanic, Italian or Irish origin. Yet there is increasing consensus among psychologists, educators and some test makers, as well as members of minority groups, that this essential condition of "sameness" is not met by the ability and achievement tests most widely used in education. These tests commonly reflect the vocabulary specializations, language styles, and cultural knowledge of English-speaking children of the majority culture. When such tests are used with children from other cultural and linguistic groups, these children are at a disadvantage. On the most simple level, we can understand that a test of history knowledge or reading skills written in English places one set of demands on children who are native English speakers and quite another set of demands on the estimated 2.5 million school-aged children in this country with limited English proficiency. But even for children who are proficient in English, discrepancies between their everyday home language and test language or between their common knowledge and the knowledge presupposed by the test are sources of unknown and uncontrolled variability in the testing situation.

Other factors varying with cultural background—motivation to perform, familiarity with test-taking, responses to stress—are known to affect performance in test-like experimental situations. In a culturally pluralistic society such as the United States, testing practices that reflect the dominant culture's standards of language function and shared knowledge and behavior provide imperfect measurement instruments for other groups. But existing tests are publicly interpreted as though they are culture fair, and low test scores are used to stigmatize minority children, to justify negative interpretations of their abilities and negative decisions about their schooling.

Current standardized tests have only limited value for holding teachers, schools, and school systems accountable for the quality of education.

The use of current standardized test to evaluate the effectiveness of education is under attack by educators and other persons concerned with education. They argue that the educational objectives tested often differ from what the school is seeking to teach. Further, it is pointed out that the tests have generally been designed to sort students for various administrative purposes rather than to determine how well they have learned what is being taught. Similar criticism is implied by the call in many states for competency standards that are enforced through new tests of performance in carrying out tasks drawn from ordinary life.

Tests exercise a limiting effect on classroom teaching.

Several national educational groups have called for a moratorium on testing. It is argued that standardized tests have no positive direct usefulness in guiding instruction, and their indirect influence—implicitly laying down goals and standards—disrupts or blocks teaching. Despite inclusion in the published tests of various subtests to identify a student's strengths and weaknesses, critics say the categories are so broadly defined, the tests are given so infrequently, and the time from test administration to report of results to teachers is so long that tests do not help teachers in their work.

Critics also find that the indirect uses of tests—their use by local educational authorities, as part of accountability procedures, to guide teachers in choice of topics and skills taught—impose undesirable limits on what teachers can do, and also on what schools and school systems can do. At the conference Ross Taylor, mathematics supervisor for the Minneapolis Public Schools, described how the Minneapolis system sought to improve its showing at the intermediate level on a mathematics achievement test, where students were several months below national norms. Analysis of the test showed that of 40 computation items, 15 were devoted to fractions. A crash program concentrating on fractions was instituted and scores went up. However, as Taylor notes, with the advent of the handheld calculator and the forthcoming change to the metric system, there are better ways to improve the mathematics curriculum at the intermediate level than signling out fractions for more intensive study.

Indirect use of tests can limit teaching in another way. The tests may be prescriptive not only of goals, but also of methods. A teacher, school, or school system seeking to build a curriculum based on discussion, primary sources for social studies, and firsthand observations for science might find

itself handicapped when it came time time for testing. There might be gains from such teaching in terms of students' feeling responsible for their own education or coming to understand how inquiry is conducted, but such gains are not likely to show up the next time a published test is administered. For immediate results on published tests, the premium approach is through use of recitation and textbooks.

Tests are too narrow in scope to provide for fair evaluation of new approaches to teaching.

Evaluation is an important part of the present effort to improve education, because without full evaluation an educational experiment loses most of its meaning. But critics maintain that the narrowness and inflexibility of published tests with regard to curriculum make them unsuitable for evaluation of new and potentially valuable approaches to teaching. In a recent review of government-sponsored evaluation, House and his colleagues (1978) question whether any battery of published tests can be used to evaluate large-scale social programs, and they add that such tests have particular drawbacks when used with young children. Looking to the future, Porter and his associates (1978) call for abandoning use in evaluation of "standardized indices" because they can be employed without any knowledge of what one is supposed to be measuring. They ask instead that the connection between an innovator's goals and the evaluator's measures be made explicit.

RECOMMENDATIONS

We have examined briefly the major elements of the current situation in testing. Educational tests are now predominantly used for four purposes. accountability, selection, evaluation, and classroom guidance. Problems in each area can be identified in the context of contemporary criticisms of tests. Selection procedures are not completely fair to minority students. The use of tests for accountability is imperfect, and now new tests for accountability are urged. The tests are not a positive force in classroom teaching and, in some regards, are perceived as inhibiting and constrictive. Finally, the tests are not broad enough in scope to allow for fully satisfactory evaluation of educational programs. These kinds of criticism of tests and testing came up repeatedly at the conference and reflect much contemporary debate outside the conference, for example at the National Conference on Achievement Testing and Basic Skills (DHEW 1979).

We now turn to the main body of this report, discussion and recommendations on the ways that the conferees believed testing, and thereby education, can be substantially improved through research and development and other related activities. *These recommendations fall into three areas. better fitting of testing to the cultural background of students, better fitting of testing to educational objectives, and combining testing and teaching.*

Better Fitting of Testing to the Cultural Background of Students

Though there have been talk about the ideal of a culture-fair test since the beginning of testing, no test has yet been constructed which meets this ideal. Early expectations that perfect "fairness" or perfect standardization could be achieve by statistical means have proved unfounded. Under pressure from minority groups, test makers more recently have made some attempts to modify the language used in test instructions and test items so that it is more representative of vernacular language varieties. While these efforts are useful, it is by no means clear that a single test instrument can be equally representative of the language patterns of all major cultural groups

in the population. Nor is it sufficient to modify surface linguistic features of tests.

Recent advances in our knowledge of language interpretation and of cognitive processes reveal that culturally different experiences and background knowledge may affect test performance in complex and subtle ways Take tests of reading skills for example. Investigators at the Center for the Study of Reading (Steffensen, Jogdeo, and Anderson 1978) in well-controlled studies, showed the importance of the match between cultural background and reading passage content on reading speed, on reading comprehension, and on retention of information. Native American Indian children and majority culture children were given two stories to read about weddings. These stories had been carefully constructed to be comparable in vocabulary level and equivalent on measures commonly used as indices of reading difficulty, but one story dealt with wedding customs of Native Americans and the other with the "typical-American" wedding ceremony as depicted in magazines. Each group showed greater reading speed, superior recall, and better performance on comprehension questions for the story in its own tradition. This outcome seems commonsensical, but its implications for testing practices are profound.

While many items in achievement tests are designed to assess how familiar a student has become with material presented in the curriculum, other items use factual material or event descriptions in order to assess skills of comprehension, reasoning, memory, or problem-solving. Unless the material used for these purposes is equally familar to all cultural groups, differences in performances are uninterpretable. The difficulties of achieving "equal familiarity" in this sense are so formidable as to make the ideal of culture-fair tests appear unrealizable and perhaps, unreasonable.

An alternative strategy, and one adopted by many investigators specializing in comparative research, is to construct different forms of a "single test"—such that the difficulty level and knowledge domains assessed remain constant across forms, but the language and illustrative material are tailored to the specific background knowledge of particular groups The aim here would be to devise "equivalent" culture-specific tests. This was, in fact, the strategy advocated by Binet, the investor of the mental test. He rejected the notion that a uniform tests could be used as a means of comparison of people from widely differing background and insisted that tests be appropriate to the background and everyday occupations of the individuals tested. Clearly the notion of perfect equivalence or comparability is also an "ideal" that can be more easily approximated for some types of tests than others. *But participants in the conference were of the opinion that recent advance in knowledge and technique make this a fruitful strategy to pursue.*

Better Fitting of Testing to Educational Objectives

Three of the major criticisms of current testing discussed above—inadequacy for accountability, negative or limiting effect on classroom teaching, and unsuitability for evaluating new approaches to teaching—derive either entirely or in considerable part from lack of fit between testing and educational objectives.

To understand how this lack of fit has come about, a look at the history of educational testing in this country is useful. The successful use of psychological and educational testing in World War I led to their wide adoption by school systems. Tests were used by the military services to select recruits for officer training and for training in various technical tasks. The method of sorting thus developed was then adapted to address the problem of sorting students in the civilian educational system. At that time, most students were not expected to finish high school and go on to college, and thus a major function of schools and colleges was to sort children and youths, encouraging only those who were most promising to go on. Much of educational testing and testing theory developed in this context of sorting. In this theory, validity is measured by correlation of the test results with some other relative measure like grades in school and college. An item is judged by its ability to spread scores out for sorting purposes rather than for its relevance to what the school is seeking to help the student learn.

In addition to these assumptions made to facilitate arranging students on a linear scale for sorting purposes, two other assumptions have tended to confuse and impede the improvement of educational testing. The first is the notion that the educational objectives of schools and colleges do not go beyond such simple skills as reading and computing and the recall of information in content areas. The second is the assumption that the attainment of important educational objectives can be adequately appraised by the use of paper-and-pencil tests alone.

We have now moved into an era where we seek to help all students achieve their full educational potential. As discussed above, we are attempting to use tests for a range of purposes much broader than sorting of students. Work on the curriculum by teachers, scholars, and textbook authors in the last two decades has made explicit a wider range of education objectives. The coming of low-cost information-handling technology makes it possible for us to escape the limitations on testing imposed by 50-year-old scoring technology based on multiple-choice paper-and-pencil tests.

Our recommendations for achieving a better fit of testing to educational objectives fall into three categories. The current movement toward criterion-referenced testing should be strengthened. The potential of the new technology should be exploited in such areas as creation of computer-based test item pools, testing of problem-solving processes, and

testing tailored to the individual. Better information must be developed and made available on tests and testing, including basic conceptions of testing, appropriate use of testing in education, and capabilities of existing tests.

Criterion-referenced Testing

A major response to the need for tests that serve purposes other than sorting has been the development of criterion-referenced testing. A criterion-referenced test determines what a student can or cannot do in a specified domain of educational objectives. Ideally, items are selected to give a proper representation of the domain. In contrast, traditional achievement tests, designed in the sorting tradition, eliminate test items that most students can answer, since these items do not produce the spread in scores desired for sorting. This latter practice tends to eliminate test items that represent what schools are trying hardest to teach and, as time goes by, may penalize better teaching by removing well-learned items in revised versions of the test. (These and other contrasts between criterion-referenced and ordinary achievement tests are discussed by Popham, 1978). Criterion-referenced tests can provide educators, parents, and others with a rather detailed picture of how well students, individually and in classes and schools, are learning in domains covered by the tests. Test makers may also develop data that allow comparison of performance in these domains among various categories of students and schools in different part of the country. For example, the National Assessment of Educational Progress utilizes criterion-referenced tests and publishes national and regional performance data that schools can compare with their own.

Test makers are increasingly producing criterion-referenced tests. *However, the theory and practice of constructing and interpreting the results of criterion-referenced tests need further development.* Preparation of good criterion-referenced tests requires more careful analysis of the content domains being tested and preparation of more tests and test items than is the case with tests designed for sorting. Some tests presented as criterion-referenced are little more than reworked versions of sorting tests, without the requisite coverage of content. Construction of criterion-referenced tests may still be strongly influenced by the traditional objective of spreading out the distribution of scores on a bell-shaped curve.

Ideally, a school system should be able to give tests well-fitted to their chosen educational objectives at various grade levels and thus determine in detail how well these objectives are being attained and how this attainment compares with that of other systems. In addition to the barriers to achieving this ideal described in the previous paragraph, most test development has

been confined to items testing simple skills or factual knowledge. The tendency to think of education in terms of the most basic skills has been intensified in recent years by the fact that some fraction of students still come out of school unable to read or to do simple arithmetic. This is a real problem for American education, but it is only one of a number of problems in a society that each year demands more highly skilled graduates from its educational system. The emphasis on our aspirations for universal literacy has tended to cause us to lose sight of broader educational objectives. Thus for example, many mathematicians, mathematics users, and mathematics educators today are concerned that the public and some educators are defining basic skills too narrowly, limiting attention to computation. Responding to this concern, the National Council of Supervisors of Mathematics (1978) published the following list of ten basic skills in mathematics:

- Problem Solving
- Applying Mathematics to Everyday Situations
- Alertness to the Reasonableness of Results
- Estimation and Approximation
- Appropriate Computational Skills
- Geometry
- Measurement
- Reading, Interpreting, and Constructing Tables, Graphs, and Charts
- Using Mathematics to Predict
- Computer Literacy

Computation is an element in all of these skills, but the formal computational skills given greatest emphasis on most mathematics achievement tests, constitute only one of the ten basic skills listed. As mathematics teachers increasingly emphasize these other basic skills, tests used to assess their success in teaching must contain an appropriate selection of items in these other areas. Similar needs exist elsewhere in the curriculum. The more broadly conceived skills being called for in mathematics and elsewhere are often ones that reflect the needs of adults in analyzing and solving practical problems that confront them in their jobs and personal lives. Similar objectives—such as the ability to solve practical problems involving computation and reading—are found in a number of the competency tests being devised by the States.

Finally we note that although there are certain things that a school will wish every student to learn, we are also interested in encouraging students to develop special interests and capabilities of their own. Expecting every student to answer every question on a test is inconsistent with this goal and

will tend to keep each student on a uniform path. A useful alternative (Zacharias*) is to make items on a test ⅓ mandatory, ⅓ choosable from a longer list, and ⅓ designed by the student to show his or her grasp of an idea of skill. As children move towards the higher grades of school their interests and skills diverge—as they properly should in a country that lists 20,000 different titles in its dictionary of occupational skills. Something other than uniformitarian testing is needed to supervise and encourage the diverse growth of children's competences.

Information-handling Technology

The fit between testing and educational objectives can be improved by taking advantage of the increasing availability of low-cost information-handling technology. We discuss three examples. computer-based item pools, computer-based testing for broader objectives, and tailored testing.

Computer-based Item Pools

The task of providing teachers, schools, and school districts with tests closely matched to their specific educational objectives can be made manageable at reasonable cost through the new technology. Central computerized pools of test items of varying complexity could be created by educational centers and test publishers. Users could be given access to these pools either through direct communication between the central computer and their local computer or indirectly through local information storage devices such as magnetic or video disks. Items would be indexed in such a way that users could assemble them to form tests suited to their needs. The system could also provide additional information on the items, including national error rates and comments by other users or critics. It would be easy to design direct access systems in which additional items as well as comments and additional error data on existing items could be entered by any user, but an arrangement in which the center acted as an intermediary would probably be better. Similarly, many teachers might desire the help of a local expert in compiling tests. Tests could be printed out for paper-and-pencil administration, but an eventual further refinement could be administration and scoring of the test via a personal computer for each student.

The statistical properties of a test as a whole are not, of course, the simple sum of the statistical properties of its individual items, and items do have interactions that are not fully foreseeable until they are tried out together in

* References given without a date refer to papers presented at the Conference (see Appendix)

The New Information-Handling Technology

We have entered a revolutionary age in the technology of information handling. The microelectronic revolution, extensively documented in the September 1977 issue of *Scientific American*, is making possible a roughly tenfold decrease each five years in the cost of the integrated circuits that are at the heart of contemporary digital computers. There are available for individual use, at relatively low prices, sophisticated calculators and computers and means of generating images on video tubes:

Handheld calculators providing the four arithmetic functions and square root now sell for about $10, and the cost of calculators providing trigonometric and logarithmic functions has declined to the same level. Calculators able to handle 400-step programs are now at the $100 level.

Handheld instructional devices providing drill and games in arithmetic ($15-25) and in spelling, with a simulated human voice ($65), are selling well.

Personal computers which include a TV tube display for letters, numbers, and graphics, keyboard, processor, a sizeable memory, and weigh less than 50 pounds, are now available for as little as $600. Accessories include word-processing devices allowing flexible construction and editing of written text and production of typed copy. These computers are comparable in capacity to computers costing hundreds of thousands of dollars a decade ago, and are only the first of many that will become available at decreasing prices or with much greater capability at the same price. Incorporation of microcomputers into TV sets could make substantial computer power available in the home to each child.

A *videodisc* system coming onto the market for $700 provides the following, in conjunction with a regular TV set:

- 54,000 separate frames of full color picture image, or alphanumeric or computer information, on one side of a disc similar in size and cost to an LP record. The 54,000 frames correspond to 30 minutes of video at 30 frames per second, any part of which can be played at regular or reduced speed.
- Random access (dialed in more expensive models) to each of the frames, which are individually numbered and can be viewed individually for any length of time desired and may contain one quarter of a page of easily legible text (or a full page with special high resolution video tubes).

Intelligent videodisc systems under development include a computer which allows controlled sequencing of frames, which may be based on student responses, as well as the possibility of transfering computer programs from the disc to the computer. Video discs can store 10 billion bits of information (*Encyclopedia Britannica* contains 2 billion bits, a human chromosome has a capacity of about 20 billion bits, and the human brain perhaps 10,000 billion). This immense storage capacity could probably include on a single disc all the computer courseware ever published.

a test, but as new tests are developed, information on tests as well as on individual items could be developed. A teacher-assembled test could not have properties identical to those of current published tests; nevertheless, with so many local variation in populations, teaching styles, and curricula, school systems might conceivably produce new tests with greater predictive validity than current tests.

In addition to increasing teachers' choices in testing, computerized item pools would help to resolve the conflict between demands for complete access to all items and the need for secrecy of items. If the pool is large, secrecy of items in the pool is unnecessary. Only the selection of items for a particular test need be secret and then only *before* that test is given. Items used once can be used again. Indeed, part of the idea of the pool is to gather statistics as well as individual comments on items.

Mathematics teachers in Minneapolis have developed a city-wide testing program for selected aspects of junior-high mathematics that is suggestive of some features of a computerized item pool (Taylor). Teachers first decide on instructional objectives, then prepare tests keyed to those objectives, and finally write generation rules for each test item. A new form of the test is generated by computer each time the test is administered. Students use old forms for practice and as a way to learn the objectives of the course. Teachers change the objectives and the tests as they gain experience and improvements occur to them.

Testing Problem Solving

Technology now provides a means to improve instruction and testing in one of the ultimate aims of education—the ability to address more complex problems of the sort encountered in work and personal life. Such problems typically require us to bring to bear a variety of things we have learned in school or elsewhere; many are actually a series of problems, where each step depends on a previous step, and where various sequences of steps can be followed, some more efficient than others. Not only a student's answer, but also the efficiency of the solution strategy are of interest. Easily-graded standardized tests for these more complex problem solving procedures have been difficult or impossible to devise, so that a very important class of educational objectives has been left untested and thus undervalued. Information-handling technology now makes such testing possible at low cost. Computer-based testing can present problems more realistically, can allow a student to proceed even when an arithmetical error is made, and can follow and evaluate the problem solving process.

This use of technology is already well developed in medical education. For example, a test of skill in diagnosis begins with a statement of the patient's complaint. The student asks questions of the patient and orders

laboratory tests. The computer then supplies the patient's answers and the laboratory results. The student can now render a diagnosis or continue to call for more information until ready to give a medical opinion. This testing procedure not only determines whether the student correctly identifies the patient's complaint, but also shows how the student goes about the diagnostic task—how he or she processes the information furnished by patient and laboratory, as revealed by the student's questions and the order in which they are asked. The results of several such sequences can help a student improve his or her diagnostic technique. This type of testing can obviously be applied in other areas. Investigation of process can be made as profound as one chooses. "Some New Things to Test For" (page 20) lists some intellectual processes used by scientists in wrestling with problems. Further examples of computerized testing are given later in this report.

Tailored Testing

In a conventional test, the test taker works through a series of items in a fixed order, marking choice of response with a No. 2 pencil. One shortcoming of this procedure is that the test cannot measure accurately the abilities of test takers at the high and low ends of the score distributions for a heterogeneous group. In tailored testing, based on computer technology, the difficulty of the items is matched to the ability of the test taker. Multiple-choice questions appear on a display panel similar to a TV screen and the test taker indicates the answer on a typewriter-like keyboard. A correct answer is followed by a harder question, an incorrect answer by an easier one. No longer need the candidate waste time on items that are too hard or too easy, and responses to less appropriate items need no longer mar measurement based on more appropriate items. As the test proceeds, each response causes the computer to revise the estimate of the test-taker's ability. When the estimate reaches a specified level of reliability, the test ends (Urry 1977).

This approach is based on a "latent trait model" developed over 15 years ago (Lord 1952) that makes it possible to give different forms of a test, all of different levels of difficulty, to subgroups of a population of candidates and to obtain comparable scores, as though all had taken the same test. Recently, investigators have been developing tailored testing procedures for use by the Civil Service Commission (Urry 1977).

Providing Better Information on Tests and Testing

An essential element in fitting testing to the educational objectives of schools and school systems is the availability of reliable information on tests and testing. Educators, parents, policymakers, and the public need to know what tests can and cannot do, how testing can affect achievement of

Some New Things to Test For

Following is a partial list of intellectual tools and strategies that physical scientists and technologists often use in their professional work (Zacharias). We should be able to test students to see if they have, or are developing facility in the use of these and similar tools and strategies.

- Asking whether the problem is one that, in principle, can be solved.
- Asking what aspects of the problem can be ignored for the time being.
- Looking for a quick guess at the answer; looking for ways to improve the guess.
- Looking for ways to break the problem down into subproblems.
- Pushing an idea to the limits; pushing it to absurdities.
- Establishing upper and lower bounds.
- Bringing in outside information to bear on the problem.
- Recognizing and exploiting symmetries of the problem.
- Changing point of view and/or frame of reference to simplify the problem.
- Shifting attention between parts and wholes to develop new approaches to the problem.
- Anthropomorphism: What would I do it I were a . . .?
- Seeking out clean approaches and solutions: taste, style, and elegance.
- Asking for help: Why, when, and of whom? Understanding that it is no disgrace to ask a question, or to be unable to answer one.
- Possessing a decent respect for multiple sources to evidence, as a basis for establishing firmness of belief.

educational objectives, and what the capabilities of specific published tests are.

Clarifying Basic Conceptions of Testing

The best known of all tests is the IQ test. The name IQ has been part of the common language for two generations or more. A person's score on an IQ test, expressed as a single number, has tended to be generally regarded as a precise, stable, objective measure of a centrally important charac-

teristic of the person. This impression has persisted despite clear statements to the contrary by leading figures in test development, e.g., Edward L. Thorndike et al. (1927) who said the following:

> Existing instruments (for measuring intellect) represent enormous improvements over what was available twenty years ago, but three fundamental defects remain. Just what they measure is not known; how far it is proper to add, subtract, multiply, divide, and compute ratios with the measures obtained is not known; just what the measures obtained signify concerning intellect is not known. We may refer to these defects in order as ambiguity in content, arbitrariness in units, and ambiguity in significance.

The common view of the significance of IQ scores has pervaded thinking about educational tests generally. The theory that has, explicitly or implicitly, dominated most thinking about educational tests is that they provide measures, analogous to length or weight, that can be used to arrange children, and what they have learned, in a simple linear order. This theory supported initial research on testing but it is obviously inadequate to our present conceptions of the complex nature of human knowledge. The present conference was not designed with the principal purpose of resolving issues in that debate. *We believe, however, that the time is ripe for a new look at these basic questions in testing. This new look must involve, in addition to leaders in testing theory, persons from various branches of cognitive science as well as teachers and scholars concerned with the substance of education. The results of this effort should lead both to new research on testing and to informational materials for teachers, teachers colleges, parent-teacher associations, and the general public, setting forth as clearly as possible what tests can and cannot do, including issues for continuing study and discussion.*

Appropriate Use of Tests in Education

Tests are more and more widely used in the educational system for accountability, selection, program evaluation, and instructional guidance. The use of tests has often become a routine bureaucratic practice to which little thought is given. Tests designed for one purpose are often used for other purposes to which they are not well suited, and critics argue that tests often have perverse effects even when used as intended. *The conference believes a thorough program of studies of the use of tests in American education is needed to provide a basis for intelligent action. The studies should document the typical uses of tests by various educational agencies for various purposes and should investigate the effects of this testing on teaching, on the education of individual students, on educational innovations, and on budgetary and other decisions, and the appropriate*

ness of these effects. These studies would make it possible to provide much firmer information on the suitability of various testing practices and how testing can be improved than is now possible.

Information on Published Tests

There are great many tests on the market today, and there is evidence that quality control is uneven. Further, because tests are kept secret from the public, there is little opportunity for individuals to challenge the findings of a low-quality test. *Our recommendation is twofold: formation of private consumer groups as sources of information about the capabilities of tests and provision for fully informative disclosure of test items.*

Consumer information groups. Private independent groups should be formed, with private support, to provide test users easy access to information about the technical properties of published tests. A single group will not be enough: there must be several that evaluate tests from their points of view rather than seek a consensus or minimal evaluation. The groups would explain from their perspectives how the tests were developed, including the domain of knowledge and skills from which the items were selected, the formulation of items and the procedures used to reduce bias. They would describe the rationale for the content and format of the tests and the procedures for scoring.

The model for such undertakings should not be the *Underwriters Laboratory*, which approves the safety of an article, but something like *Consumer Reports*, which gives broadly accessible information about the factors of quality in an article and which then gives item-by-item information about how articles measure with regard to those factors. We are *not* recommending that tests be rank-ordered, or given ratings like "Acceptable," "Best Buy," etc. Currently, the Buros *Mental Measurement Yearbook* provides reviews of recently published tests, but these are given in a lengthy and detailed form that is probably not ideal for the unsophisticated consumer who wants to review what is available before choosing a test. We are not recommending that this activity be initiated by the federal government or a sponsored subsidiary such as an educational laboratory or center. Federal quality control over tests could evolve towards a national curriculum and in any event would be less desirable than several sources of test information developed through private initiative.

Fully informative disclosure of test items. Students and parents affected by decisions based on tests should be able to see the tests—that is, the individual items and the student's answers—on which these decisions were based. This would seem at once a fundamental human right and a necessity, since testing procedures are fallible. The research community and any concerned layperson should have easy access to test items and the

grouped responses of test takers so as to form independent judgements about the quality of the items and the presence of possible bias, and freedom to publish the items on which their judgements are based. Arguments in a debate about an item or a test would seem to rest ultimately on appeal to items themselves, and without access to the items one cannot participate fully in the debate.

In considering the argument for secrecy, a distinction should be drawn between secrecy before administration of the test and secrecy afterward Before administration, the topics to be covered may be revealed, but not the specific items and their answers. After administration, why should any matter be kept secret? Test publishers maintain a policy of secrecy after the test because the need constantly to redevelop old tests that had become public knowledge would increase the costs of test construction. One exception to the rule of secrecy is a federally-supported program, the National Assessment of Educational Progress, which releases 40 percent of its questions after each round of testing. Of course, most companies do release sample items which show some aspects of the questions asked. The computerized item pool described above requires no secrecy about items available for use on tests. We recommend also that the consumer information groups address themselves to the question of fully informative disclosure of test items.

Combining Testing with Teaching

Instructional guidance is the educational activity which is least served by existing tests. Yet the interaction between teacher and pupil is at the heart of schooling. Further, use of tests for purposes outside the classroom—accountability, selection, evaluation—should come out of classroom process, not be imposed on it like a foreign body. *Conferees envisioned, as a future ideal, testing merged into the teaching process and providing timely and rich feedback to the student, the teacher, and other interested parties as desired. The recommendations envision sustained development and research over the next decade or more exploring the use of testing in instructional guidance. Some approaches to this ideal already exist, but they are seen as only a beginning to be developed further and supplemented by other approaches.*

We see four elements as central to the development of this new combination of testing with teaching. full use of what cognitive science can tell us about the processes involved in learning and performing the intellectual tasks to be tested, exploitation of the new information-handling technology, strong participation of teachers and scholars in the subject areas, and good fit with the natural classroom situation. The

contributions of each will be discussed separately, with the understanding that they must be brought together in the design of new teaching-testing systems.

Cognitive Science

Cognitive science today is a rapidly-growing interdisciplinary movement including psychologists, computer scientists, philosophers, linguists, anthropologists and educators all converging on the analysis of human intellectual processes. An important component of this effort has been the detailed study of learning by students at all educational levels from preschool to college, coupled with a systematic effort to model partial and progressive states of knowledge using the computer. Any attempt to teach a subject involves some kind of theory or assumptions about what high-level performance in the area is like, as well as an explicit or implicit theory of the learning process. *Contemporary cognitive science is conducting a much more searching appraisal of both performance and learning than has heretofore been possible. The work in cognitive science is far more theoretically compelling than the "learning theories" of twenty years ago and, at the same time, much more closely tied to direct studies of learning in school environments. Development of better systems of instructional guidance will be greatly aided by this knowledge of what is involved in the student's progressive encounters with subject matter.* Presentations at the conference discussed efforts dealing with mathematics, reading, and writing, some of which will be described here.

One of the attributes of a good teacher is the ability to diagnose underlying misconceptions from a student's answers to a set of problems. The BUGGY computer program (Brown and Burton) can undertake this task for certain aspects of mathematics instruction. The computer program includes "correct models," which represent the various ways to obtain correct answers, and "diagnostic models," which represent various ways that students typically obtain wrong answers. Like a good teacher, BUGGY is not limited in instructional guidance simply to indicating which answers are correct and which wrong, but can also indicate which of many misconceptions a student may harbor: the "bugs" in his procedures. There are many possible bugs in children's arithmetic (e.g., always subtracting the smaller digit from the larger: $1928 - 573 = 1455$). Brown and Burton are devising a computer program allowing diagnosis of more than one bug, a task generally beyond the capability of a teacher.

It is tempting to assume that students make mistakes because they do not follow procedures very well: that the primary cause of error is simply inability to carry through a sequence of steps properly. But good teachers operate on the assumption, and the BUGGY program begins to

demonstrate, as Brown and Burton note, that students are remarkably competent followers of procedures; the difficulty is that they often follow the wrong procedure. That is why cognitive science can aid good teaching by investigating not only the student's answer to a set of problems, but also the processes by which he or she obtained those answers, so that the correct processes can be learned.

Investigators are also developing psychological models of how children learn to read and write, based in part on comparative studies of how beginners and accomplished performers proceed. As described in a presentation at the conference (Collins), reading is a constant process of forming and correcting hypotheses, of which the reader is ordinarily unaware. Children sometimes fail to understand a passage not because they lack "reading ability," but because they lack some piece of knowledge needed to develop the correct hypothesis. Experienced readers face similar difficulties fixing the meaning of passages when they lack the necessary background knowledge or do not know the context in which a sentence is set. To give two extreme examples, if a reader is to understand the sentences,

"the notes were sour because the seams were split" and
"the house blew it",

the reader must know that one sentence concerns bagpipes and the other gambling activities.

Reading comprehension is necessarily intertwined with background knowledge. a passage must be about something. Students from different homes come to the reading task with differing background knowledge It is possible to make progress in testing reading comprehension, rather than background knowledge, by relating the test to matters that have been specifically taught. Cognitive psychologists believe they can develop diagnostic procedures for aspects of reading analogous to those in the BUGGY program for arithmetic.

The Degrees of Reading Power test (Koslin et al) is based on recent work in psycholinguistics. The fluent reader is seen as using syntactic and semantic cues from the text, along with prior knowledge of language and content, to reduce uncertainties and confirm or disconfirm predictions concerning the meaning. A student can be considered to comprehend when he or she has eliminated all but the correct meaning of the text. The test items are passages of varying difficulty in which certain words have been deleted from the text. Subjects are asked to select the word deleted, from a list including several alternatives. Having carried out the comprehension process described above is a necessary and sufficient condition for choosing the right word. Persons taking the test can be allowed to inspect a list of topics in the item bank so that a reading test comprising items about which

the subject knows something can be assembled, and possible effects of differing background knowledge can thus be reduced.

Interactive Teaching-Testing and Technology

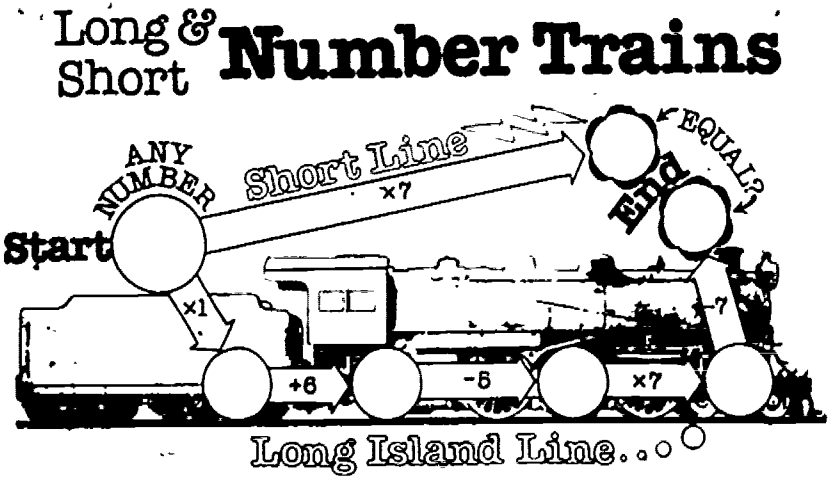
Generally, human beings learn best when there is rich and immediate response to the learner's initiatives and when there is the possibility of branching, exploratory interplay between the learner and the teacher, and testing is part of the interplay. Most schools are not able to provide that kind of learning environment because it would require something close to a one-teacher-one-student relationship. The rapidly increasing capacity and decreasing cost of electronic information-handling technology make interactive teaching-testing arrangements both feasible and economically attractive.

We have already described BUGGY, which represents an advance in cognitive science and also an exploratory effort in interactive teaching-testing, as is the example from medical education presented earlier. A number of other examples are provided by the PLATO system of computer assisted instruction, which can provide graphic displays as well as words and numbers on a screen. A number of games devised for PLATO require solving mathematical problems. In *How the West Was Won*, a player is given three digits and must create a single integer using all three digits once and only once and using parentheses and the arithmetical operations +, −, and x not more than once each. Thus 4, 1, and 6 could be used to produce $(4+1)x6=30$, or $6+1-4=3$ among other integers. The integer chosen (and verified by the computer as correctly generated) moves the player's stagecoach or train as in a board game. Since moves have different consequences, the player should seek the most advantageous move by exploring the several integers that can be generated from the given digits and the consequences of the resulting moves. The student, playing against a fellow student or the computer, is thus exercised in both arithmetic and game strategy. Burton and Brown (1979) have devised a computer program that characterizes and evaluates a student's strategies in playing *How the West Was Won* and in other activities, proving a basis for improving them. Other PLATO arithmetic programs include two units on fractions, one involving the cutting of pizza pies and the other estimating the fraction corresponding to a point on the number line marked by a balloon which bursts when the estimate is a good one. These and oher PLATO programs in arithmetic, in which a student is constantly tested in a non-threatening way, have proved to be engrossing and effective supplements to regular instruction.

Conferees proposed exploration of a new vision of a learning and testing environment extending what has been just described. In this vision

of the future, school tests as we know them would cease to exist. The instrusive, specialized, institutionalized activity called testing would be absorbed into a new kind of learning and testing environment. Computers could accept inputs from students and teachers on an almost continual basis, extracted from the rich tapestry of ongoing learning activities Instructional systems would accumulate an educational portfolio for each student, including a wide range of interrelated performance and situational descriptions. One would be as unlikely to cease all instructional activities in order to test a student as one is to stop conversing with a child in order to test his or her linguistic competence. Instead, testing would be a particular aggregation and analysis procedure applied to a continuously collected data base. Some of these aggregations would have an immediate impact on ongoing learning activities, others would be remote from the moment of data collection.

Interactive teaching-testing materials for schools can also be developed without a computer. As part of its system of instruction in computation and measurement for elementary schools, Project TORQUE has producted paper-and-pencil tests that provide immediate feedback to student and teacher. One type of test used is exemplified by the following figure.



TORQUE

Students choose a number and carry out the two branching sequences of operations. The test is self-checking. if the two sequences do not yield the same number, students know they must recalculate. These tests can. be made public because the actual problem is not fixed until the starting number is selected. The same test can be used over again with a different

number, selected by the teacher if desired. These self-checking tests have broader implications as well. Some students will be curious about why the tests work. What is it about the different paths followed that produces the same answer? Finding the reasons constitutes an introduction to algebra. These and other TORQUE tests can be categorized in terms of the specific skills which they assess. Consequently, they can be useful in demonstrating that students have mastered certain skills if required by school districts or state agencies.

Subject Matter

To construct good tests to assess what a student knows and can do in a domain of knowledge requires not only skilled test developers, but also outstanding teachers and scholars experienced in the content domain, as part of the testmaking team. Persons with such background who look through collections of published tests in their area typically find in many of the tests numerous items that they regard as inappropriate exemplars of the knowledge area, as seriously ambiguous in wording, and too often as simply wrong. This leads them to conclude that knowledgeable persons were not involved in creation of the tests in question. Involvement of knowledgeable teachers and scholars in testmaking, important for any test, becomes absolutely essential in the development of testing-teaching systems and tests of problem-solving processes. The conference heard papers from teachers and scholars in the sciences, mathematics, and reading and writing which discussed educational objectives and intellectual processes in their areas with a richness and depth that only persons with their knowledge can bring to the testmaking process. Rough examples to be found in the present report include the outline of new things to test for on page 20 and the discussion of bugs in arithmetic processes on page 24.

The Natural Classroom Situation

The teaching-testing systems envisioned in this report are intended as aids to the teacher, rather than as substitutes for the teacher. Systems developed on this basis must take account of the natural classroom situation or be in peril of being rejected or poorly used. Developers should make full use of what is already known about the way the instruction and instructional guidance take place in tyical classrooms at the grade level and in the subject area in question and about the successes and failures of other newly-introduced systems. However, the new system envisioned is significantly different from what has gone before, and its reception by teachers and students is not predictable in detail. Research on these matters must be built into development projects and be used to improve the system.

The number of investigators interested in direct, intensive observation

of classrooms is growing. They include anthropologists, psychologists, sociologists, and teachers with a talent for stepping back from their experiences and describing them. Conference participants working in the area provided several examples of how tests could be made more useful in the classroom. Test results closely related to the day's activities rather than to larger units are likely to be most useful to both the teacher and the student, and feedback on what probably went wrong would be even more useful. However, the need is not uniform across students. Teachers often believe they have enough information about many students in the classroom. For some students the teacher may feel the need for a great deal more information. The teacher doesn't understand the student, has "tried everything and doesn't know what to do." Testing procedures providing insight into such students' intellectual functioning could be very helpful to teachers. Both examples illustrate the more general principle that teachers, like other people, prefer arrangements that adapt to their needs rather than requiring them to adapt themselves to a rigid system. We do not wish to imply, however, that new systems should respond only to needs currently perceived by teachers. The most useful innovations may be those undreamed of prior to their invention, but they too must be tried in real situations and modified as necessary.

Conclusion

How are we to pursue this vision of testing merged into a teaching-testing system, fitted to the natural classroom situation, drawing upon cognitive scientists and teachers and scholars in the subject areas, and exploiting the rapidly developing information-handling technology? One way is to continue and perhaps expand support for research on classroom processes and human cognition and for development of new technologically-based testing and testing involving persons from subject areas. Eventually, however, these points of view must be brought together if we are to have the working testing-learning systems envisioned at the conference. Project TORQUE, with support from private foundations, is developing one such system in computation and estimation for schools, and the School District of Philadelphia (Charp), with support from local and Federal sources, has developed a different system based on behavioral objectives in various areas of education. We need to develop and experiment with more models. Projects to develop these models require research components if they are to achieve their full potential. Furthermore, development projects are often excellent sites for fundamental research in such areas as classroom processes and human cognition.

REFERENCES

Burros, Oscar K. (1978) *The Eighth Mental Measurements Yearbook,* 2 vols., Edison, N.J.: Griffon Press.

Burton, Richard R. and Brown, John Seely (1979) An investigation of computer coaching for informal learning activities. *International Journal of Man-Machine Studies* 11:5-24.

Department of Health, Education, and Welfare, National Institute of Education (1979) *Achievement Testing and Basic Skills. Conference Proceeding.* Washington, D.C.: Government Printing Office.

House, Ernest R., Glass, Gene V., McLean, Leslie, and Walker, Decker F. (1978) No simple answer. critique of the "Follow-Through" evaluation. *Harvard Educational Review* 48:128-160.

Lord, F.M. (1953) A theory of test scores. *Psychometric Monograph* No. 7.

National Council of Supervisors of Mathematics (1978). Position paper on basic mathematical skills. *The Mathematics Teacher.* 71:147-152.

Popham, W. James (1978) The case for criterion-referenced measurements. *Educational Researcher* 7:6-10.

Porter, Andrew C., Schmidt, William H., Floden, Robert E., and Freeman, Donald J. (1978) Practical significance in program evaluation. *American Educational Research Journal* 15:529-539.

Steffensen, M.S., Jogdeo, C., and Anderson, R.D. (1978) A cross- cultural perspective on reading comprehension. Technical Report No. 97. Champaign, Ill.: Center for the Study of Reading.

Thorndike, Edward L. et al. (1927) *The Measurement of Intelligence.* New York: Teachers College Bureau of Publications.

Urry, V.W. (1977) Tailored testing. a successful application of latent trait theory. *Journal of Educational Measurement.* 14:181-196.

NOTE. References in the text given without date refer to papers presented at the Conference (see Appendix).

INDIVIDUAL COMMENTS

Two participants asked that the comments below be appended to this Report.

Comments by Donald Ross Green, CTB/McGraw Hill:

The report begins by noting that current standardized achievement tests are widely used and that they must therefore be considered useful by many people. The report then proceeds to make criticisms of these tests both in the section so labeled and thereafter without noting either their merits or that many people do not accept these criticisms as reasonable or as based on facts. Perhaps the merits of these tests are indeed so evident they need not be reiterated but certainly the alternate points of view about at least a few of these criticisms should be cited. This commentary will therefore give a brief rebuttal to three of the criticisms and then try to reinforce the most cogent point in the report.

Curriculum Fit

A truly difficult issue is the matter of the fit between the curriculum and content of the test. There are substantial differences among the major achievement test batteries in what skills are emphasized and at what grade levels these skills are tested. Thus school systems typically have a test selection committee which includes classroom teachers and curriculum supervisors. These committees pick the tests they believe best, usually emphasizing the fit of the test to the curriculum of the system. (The Center for the Study of Evaluation at UCLA has prepared forms and procedures to assist people in studying tests this way.[1]) However, it is common to find people in a school system who disagree with the curriculum specifications of their system and therefore with the test selected. These people and others

[1] Hoepfner, R. *Achievement test selection for program evaluation.* In Wargo, M. & Green, D.R. (Eds.), *Achievement testing in minority and disadvantaged students for educational program evaluation.* Monterey, CA: CTB/McGraw-Hill, 1978.

often do not follow the curriculum specifications closely in either program planning or instruction. Partly for this reason and partly because of the vested interest of the larger state and national communities in the outcomes of education, some evaluation experts do not endorse program evaluations which use tests specifically designed just to fit a particular local curriculum. They argue that a good evaluation is broad and tells one about student learning in the areas not emphasized as well as in those that are.

Bias

The most emotional issue of testing concerns racial and ethnic bias, especially in the context of selection. Although the conference was concerned with testing of achievement not aptitude, most of the statements about bias refer to selection and do not necessarily apply to achievement tests. The evidence about achievement test bias does suggest that there is some bias in standardized tests, that this amount is probably small in most skill areas for most minority students, and that there is probably more bias in unstandardized measures. Almost all the currently used standardized achievement batteries have been through editorial and empirical study leading to revisions designed to reduce this bias. Nonstandardized tests lack the data, and other methods of assessment lack the objectivity that permit one to study this matter, facts which preclude any serious effort to either identify or reduce such bias

Since education is a process of enculturation, a prerequisite for the complete elimination of bias is a truly multicultural curriculum and program. However, this is a distant goal and as long as there are cultural differences among groups in this society that are not fully understood or accepted by all concerned, some cultural bias in the measurement of achievement will occur. In some instances the bias will be against the majority group but logically it will more often be against minorities.

Test Security

It is averred that standardized tests are kept secret. The basis for this assertion seems to be the fact that school systems do not like to let students study the specific answers to the questions on a test ahead of time. This position is reasonable and does not preclude student practice on similar test items ahead of time. In fact all published tests provide some practice as part of the test. However, it may be desirable to provide unskillful students with substantial additional opportunities to practice taking test items of the sort to be given. Many teachers do this and some school systems do this regularly.

School systems who buy achievement tests are free to distribute copies of the test before or after its administation to teachers or others as they see

fit. In fact many standardized test reports provide information about student responses item by item. These data can be reviewed with students and/or parents in conjunction with the item descriptions usually provided or with the test booklet itself. Most teachers appear to prefer the first of these two procedures and so do most school administrators.

Frequently evaluation designs call for both fall and spring testing with the same test because this procedure provides the most accurate measure of growth. Item by item reviews with students in the fall are clearly inappropriate here as is also true when the school intends to use the same test booklets the following year.

Interactive Teaching-Testing and Technology

Finally this report properly asserts that there is limited value to the traditional norm referenced achievement test for the guidance of classroom instruction. Not only are they given only once (occasionally twice) a year but they are not designed to perform this task as their principal function Therefore aside from providing useful feedback about progress on a year-to-year basis they are of relatively little help to teachers and students To the degree these tests provide detailed criterion referenced information (which some of them do) they may also be useful for planning classroom instruction; but by and large they are much more useful to curriculum planners, school administrators, and others who have to make school or system-wide decisions.

Better procedures for helping teachers plan and execute instruction have long been needed, although the many criterion referenced tests now in use are a step in the right direction. I support the recommendations of this committee concerning the use of our growing knowledge of cognitive processes and instruction in conjunction with computer technology to provide ongoing instructional guidance to students. One caveat is needed: school learning is primarily a social process and any machine-oriented procedures which do not fully recognize this will prove ineffectual

I also would note pessimism about any short term (e.g., less than ten years) possibility of using the data so collected to replace that now provided by standardized tests for assessing group status and large-scale program evaluation. Reasonable procedures for making use of these data do not exist.

Comments by Jerrold R. Zacharias:

I. The report called "Testing, Teaching and Learning" suggests a consensus when indeed there was none.

II. a. Some of the participants believe that meaningful numerical scores cannot result from any tests, good or bad; cannot be applied to people of any sort, for any purpose or by any means whatsoever.

b. Some participants believe the opposite of this and some waffle as did the report.

III. Some of the participants believe that cognitive science, except for the wisdom of the ages, is not yet in shape to be applied to schools. The report says the reverse.

IV. Everybody belongs to some sort of minority group, with special abilities, disabilities and interests. The report refers to the present standard "Minority Groups".

V. The emphasis on the various roles for computers in the suggested "Computer-based Item Pools" results in overlooking the hard work by sophisticated people needed to make and to test test items and tests.

VI. Inadequate reference is made to the present scandalous inadequacy of the subject matter which passes for mathematics and language learning.

VII. No decent research on tests can take place if it depends on commercial tests as they now exist. This leaves NIE with the need to violate its own mandates before research gets started. No tickee, no research.

APPENDIX

Subjects of Papers Presented at the Conference

I. INTRODUCTION

Educational Objectives and Educational Testing
RALPH W. TYLER

Technology
ARTHUR S. MELMED

II. EDUCATIONAL OBJECTIVES AND COGNITIVE MODELS

The Sciences and Technologies
JERROLD R. ZACHARIAS

Mathematics
ROBERT B. DAVIS
LEON HENKIN

Mathematics: A View from the Schools
ROSS TAYLOR

Information Processing Analyses of Mathematical Problem Solving
JOAN I. HELLER and JAMES G. GREENO

Reading
ALLAN COLLINS and SUSAN E. HAVILAND

Writing
CARL BEREITER

A School Perspective on Language
PARKER DAMON

Automated Dictionaries
GEORGE A. MILLER

III. TESTING

Some Emerging Trends in Testing
NORMAN FREDERIKSEN

Cultural Considerations: African-American
ASA G. HILLIARD, III

Cultural Considerations: Hispanic-American
AMADO M. PADILLA

Instruction and Testing in the Philadelphia Public Schools
SYLVIA CHARP

Project TORQUE
JUDAH L. SCHWARTZ and EDWIN F. TAYLOR

Diagnostic Models in Basic Mathematical Skills
JOHN SEELY BROWN and RICHARD R. BURTON

Effectiveness Measure in Reading
BERTRAM L. KOSLIN, SANDRA KOSLIN and SUSAN ZENO

Note. These papers, together with the present Chairmen's Report, the reports of the working committees at the conference, and the report of the National Conference on Achievement Testing and Basic Skills are published in a separate volume *Testing, Teaching and Learning. Report of a Conference on Research on Testing,* which is for sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.

☆ U. S. GOVERNMENT PRINTING OFFICE 1979-631-367/2858